CSE 601: Data Mining and Bioinformatics (Fall 2014)

Homework 3: Clustering Analysis for Complex Networks

Team Members:
Neeti Narayan – 5009 8029
Amitha Narasimha Murthy – 5009 8028
Vivekanandh Vel Rathinam – 5009 8075

Objective:

The objective of this assignment is to implement the Markov Clustering Algorithm and apply the algorithm to the given three datasets, AT&T Web network, physics collaboration network, and the yeast metabolic network

Introduction:

The MCL algorithm is a fast and scalable unsupervised cluster algorithm for graphs. This algorithm draws intuition from random walks and it is based on the markov property

"At each step the system may change its state from the current state to another state, or remain in the same state, according to transition probabilities."

Clusters in a graph have the property of having many higher length paths between nodes in a cluster than between nodes in different clusters.

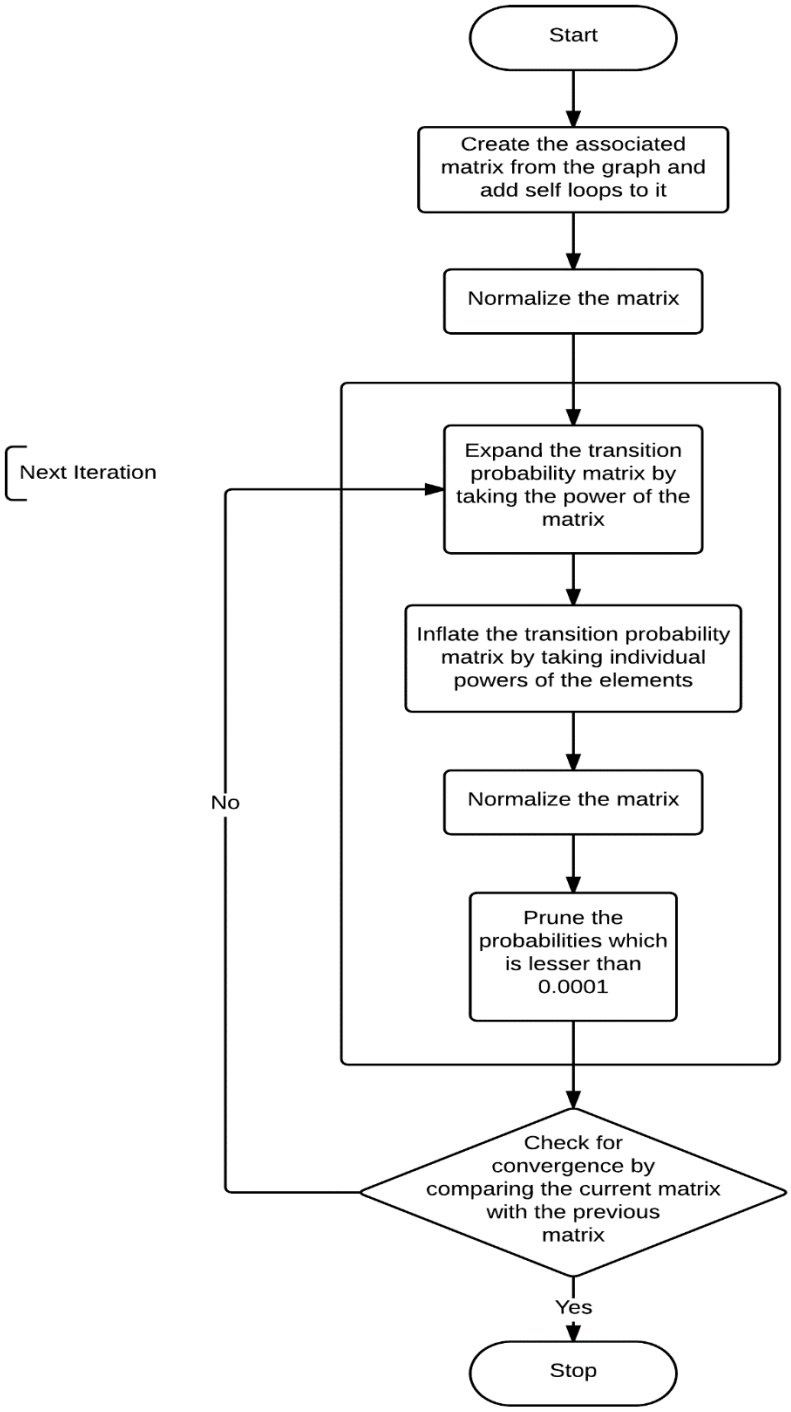
The MCL algorithm uses inflation and expansion to transform the transition probabilities which results in the above characteristic. Inflation to strengthen the transition probabilities of strong neighbors and weaken it for less popular neighbors.

Implementation of the algorithm:

We have implemented the MCL using python as the programming language. We give as input to the program the following

1. Path of the file which has the vertices and edges `[-f filepath]`
2. Expansion factor(e) `[-e expansion]`
3. Inflation factor(r) `[-r inflation]`

IMPLEMENTATION OF MARKOV CLUSTERING ALGORITHM



ANALYSIS OF MCL ON VARIOUS DATASETS:

1) AT&T Web Network

The expansion and inflation factors were plugged in differently to find the best possible set of clusters.

Expansion Factor	Inflation Factor	Clusters
2	2	55
2	1.75	13
2	1.5	7
2	**1.37**	**5**
2	1.25	1
3	2	8
3	1.75	6
3	1.65	5
3	1.625	4
3	1.5	2

We found out that two optimal clusters were formed when compared to the ground truth cluster

a) Expansion factor = 2, Inflation factor = 1.37 – Figure 1.1
b) Expansion factor = 3, Inflation factor = 1.65 – Figure 1.2

Out of the two optimal clusters, the clusters from Figure 1.1 were properly seperated than the one formed in Figure 1.2

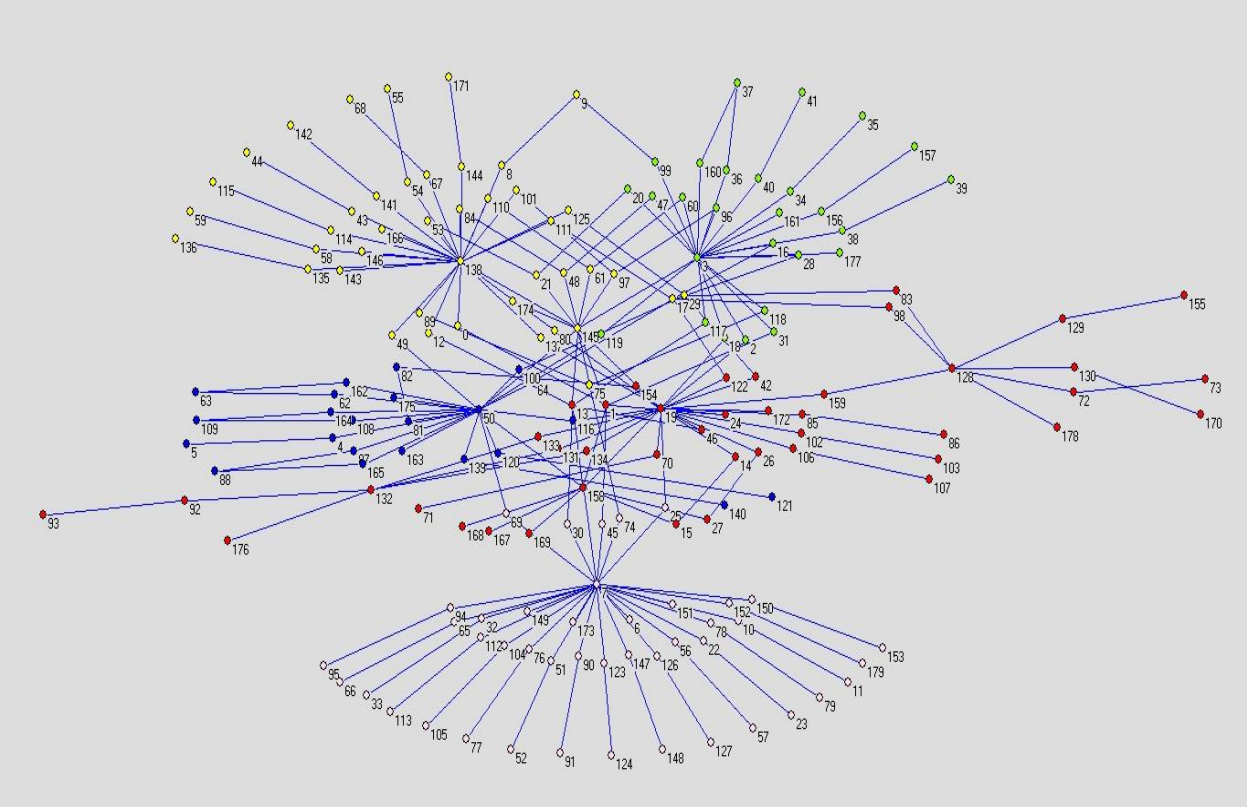


Figure 1.1 : Clusters for ATT Web (Expansion Factor: 2, Inflation Factor: 1.37)

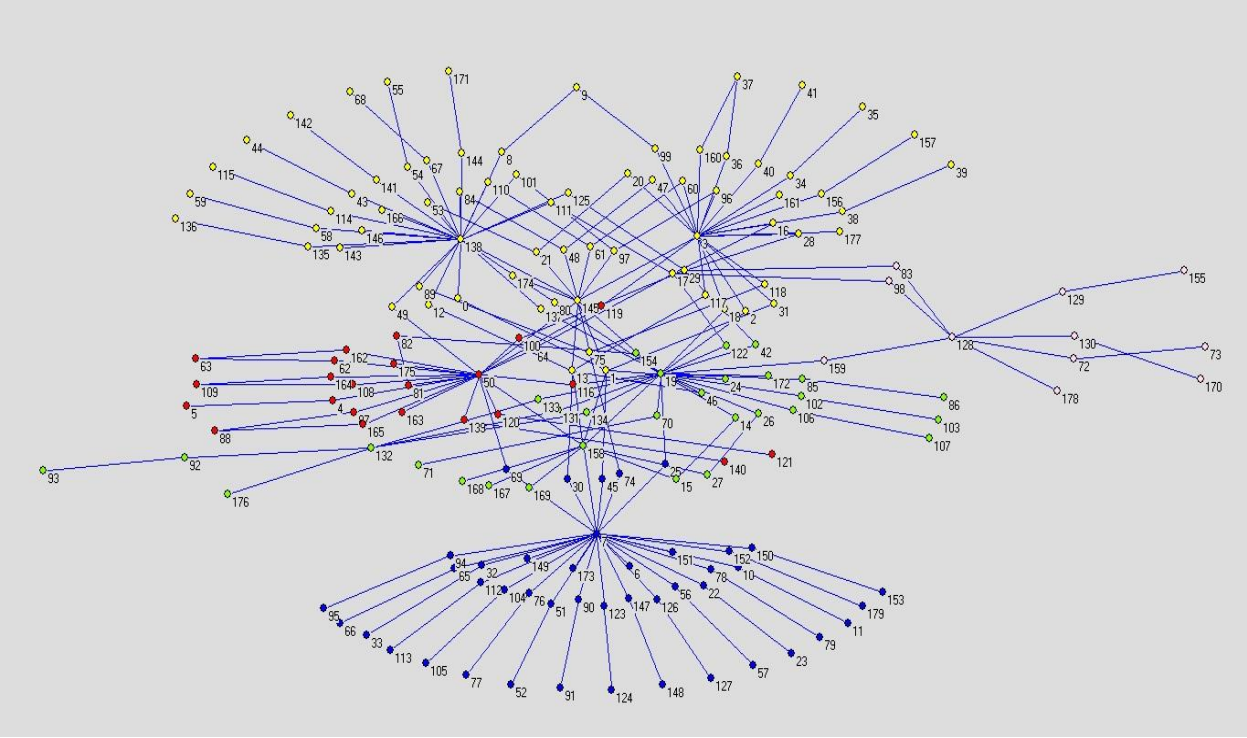


Figure 1.2: Clusters for ATT Web (Expansion Factor: 3, Inflation Factor: 1.65)

2) Physics Collaboration Network

The expansion and inflation factors were plugged in differently to find the best possible set of clusters.

Expansion Factor	Inflation Factor	Clusters
2	2	24
2	1.75	21
2	1.5	14
2	**1.27**	**5**
2	1.25	4
3	2	14
3	1.75	10
3	1.5	6
3	1.45	5
3	1.25	1

We found out that two optimal clusters were formed when compared to the ground truth cluster

a) Expansion factor = 2, Inflation factor = 1.27 – Figure 2.1
b) Expansion factor = 3, Inflation factor = 1.45 – Figure 2.2

Out of the two optimal clusters, the clusters from Figure 2.1 were properly seperated than the one formed in Figure 2.2

For instance, in figure 2.1, the nodes in the center and the nodes in the right side are separated into two clusters, whereas in figure 2.2, both are merged into the same cluster.

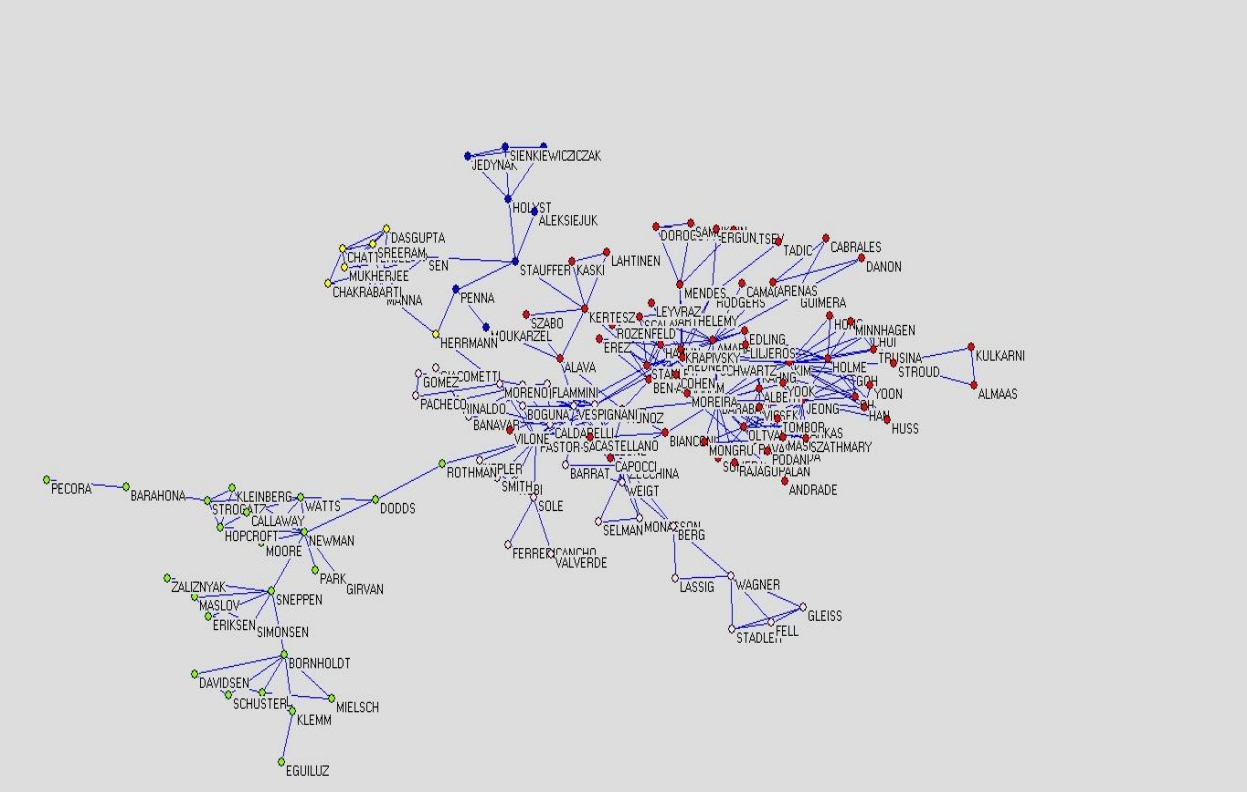


Figure 2.1: Clusters for Physics Collaboration (Expansion: 2, Inflation: 1.27)

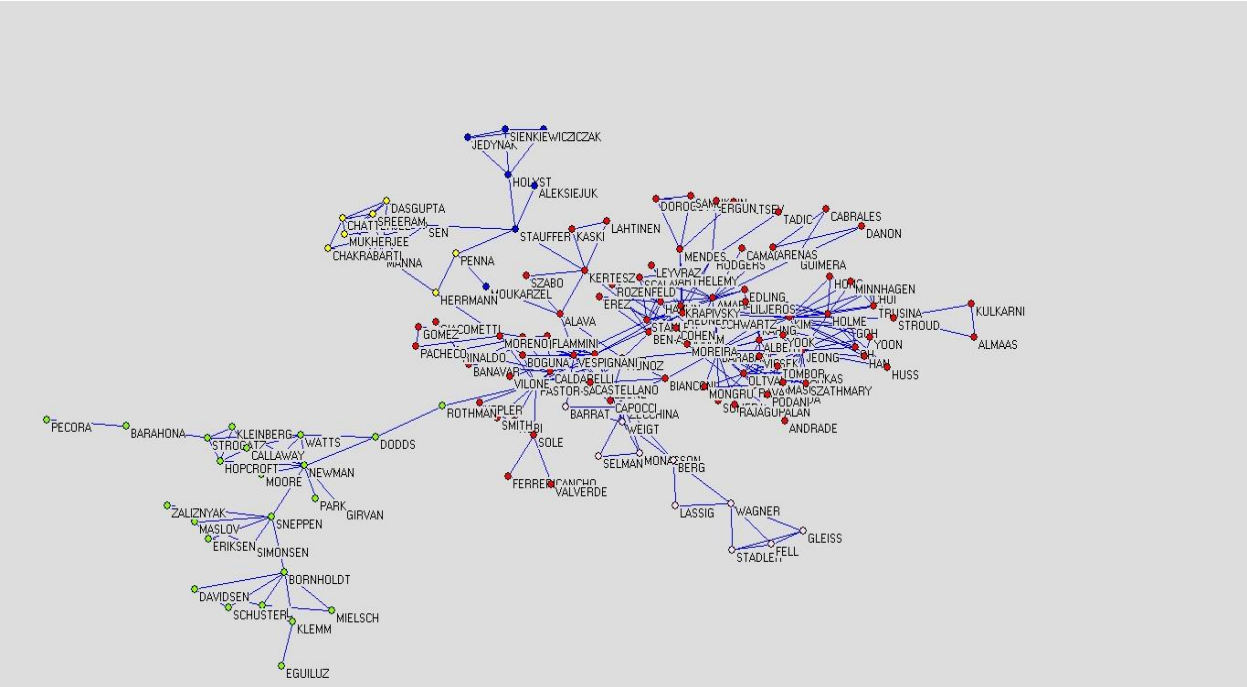


Figure 2.2: Clusters for Physics Collaboration (Expansion: 3, Inflation: 1.45)

3) Yeast Metabolic Network

The expansion and inflation factors were plugged in differently to find the best possible set of clusters.

Expansion Factor	Inflation Factor	Clusters
3	2	53
3	1.75	43
3	1.5	18
3	**1.34**	**8**
3	1.25	1
4	2	38
4	1.75	22
4	1.5	9
4	1.46	8
4	1.25	1

We found out that two optimal clusters were formed when compared to the ground truth cluster

a) Expansion factor = 3, Inflation factor = 1.34 – Figure 3.1
b) Expansion factor = 4, Inflation factor = 1.46 – Figure 3.2

As you can see below, both figure 3.1 and figure 3.2 are well performed set of clusters.

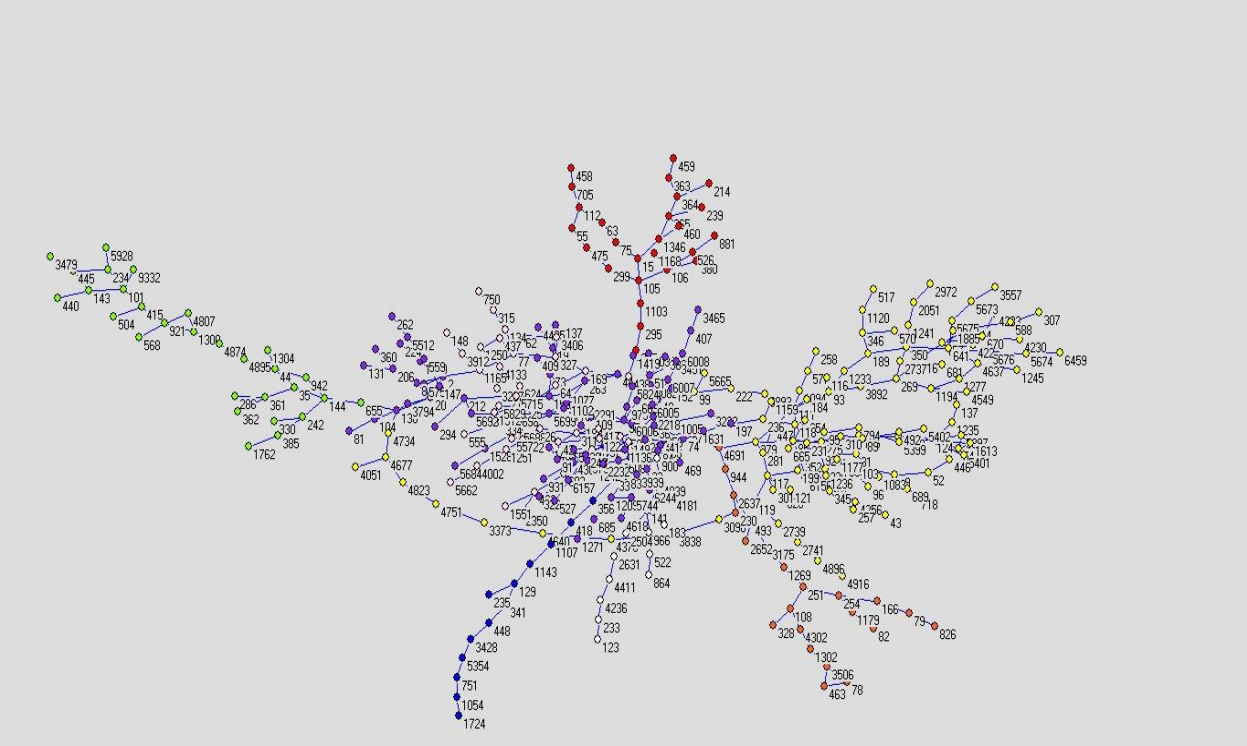


Figure 3.1: Clusters for Yeast Metabolic network (Expansion: 3, Inflation: 1.34)

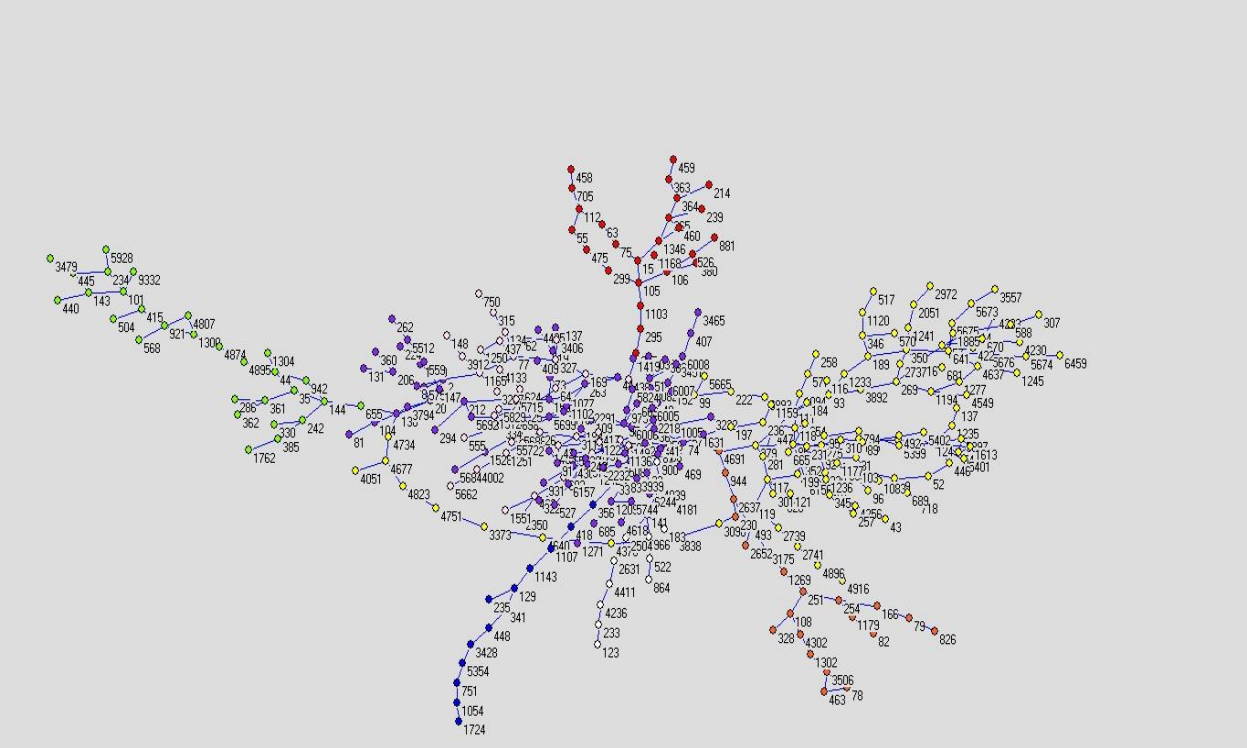


Figure 3.2: Clusters for Physics Collaboration (Expansion: 4, Inflation: 1.46)

Conclusion

Thus, the Markov Clustering Algorithm has been implemented and applied on the given three datasets and the clustering results were visualized using Pajek.

References

1) http://micans.org/mcl/
2) http://www.cse.buffalo.edu/faculty/azhang/cse601/Markov-Clustering-Algorithm.ppt